Analysing Coffee Drinkers Preferences and Habits

CMPT 353

Daniel Dawda

301445405

The Purpose

In this project I will be assisting a local business owner by providing them with useful information in order to run a successful cafe. To do this I will focus on identifying who the biggest coffee buyers are, what types of drinks are the most popular, and what flavours/coffee beans people prefer. Finally I will develop a machine learning model to predict which type of individual will pay the most amount of money for coffee, which will help the cafe determine which features (such as age, gender, etc.) lead to the greatest amount of spending on coffee so that they can attract these demographics accordingly.

The Data

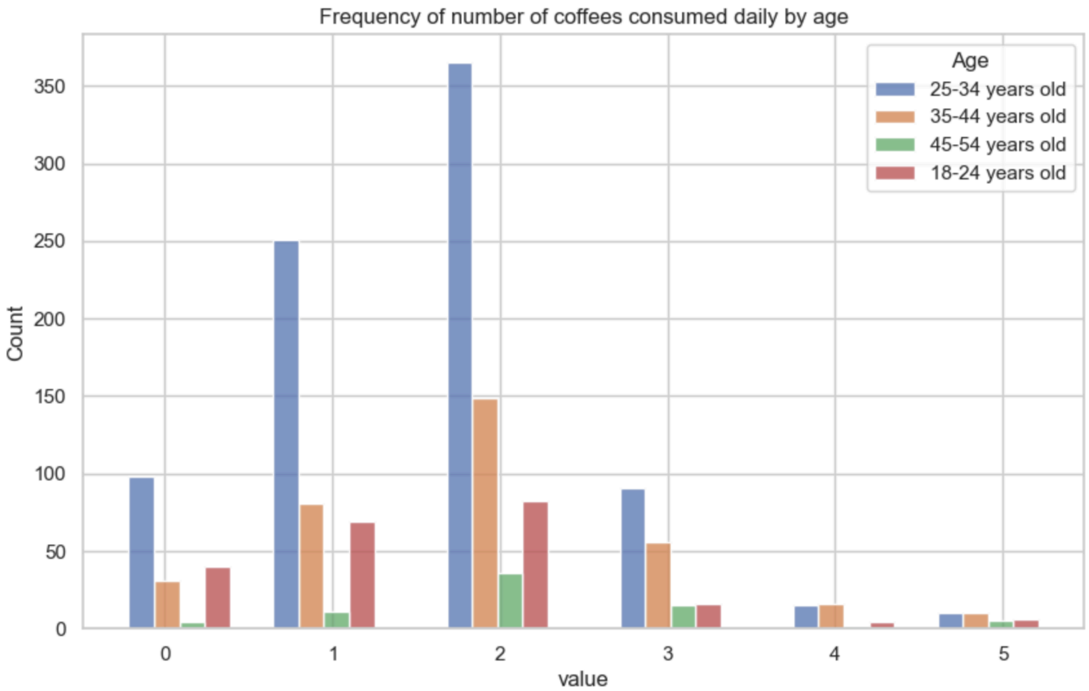
The data I will be using for my analysis is James Hoffman's Great American Taste Test. This data comes from a survey given by coffee youtuber James Hoffman to his fans and anyone else who participated in the taste test. There were 4043 respondents and many different columns I could use for my analysis. Cleaning the data was mostly renaming annoyingly named columns, mapping certain values to other values, changing data types, adding columns to identify espresso drinks and high spending individuals, and selecting only the columns I need.

Techniques

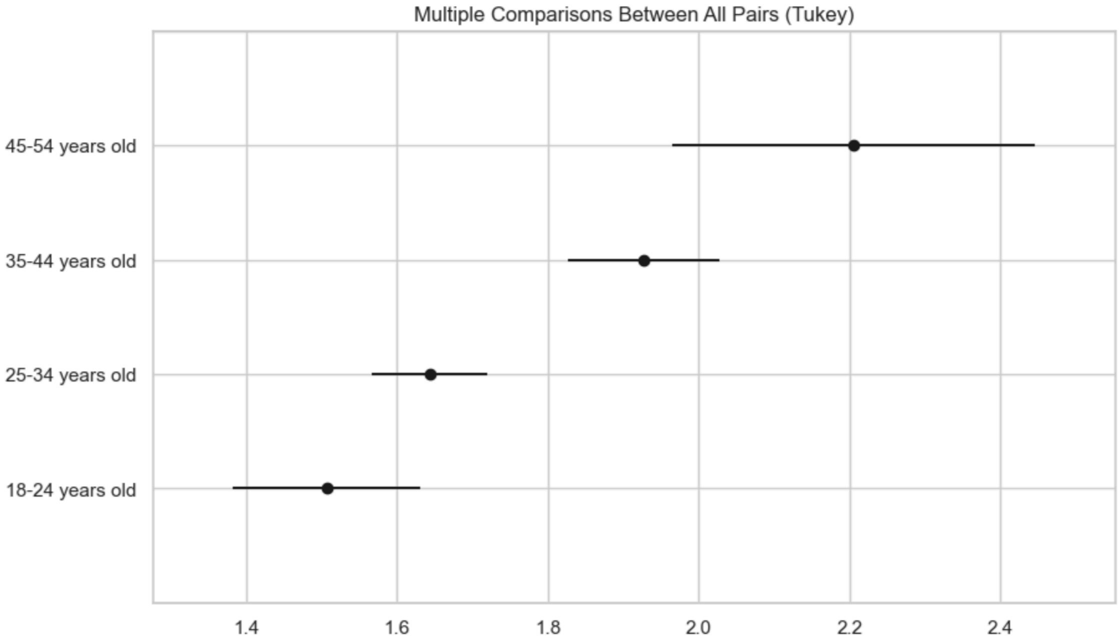
In my analysis (notebooks provided in github) I deployed many techniques that we covered throughout the course. To list off the techniques I used: ETL process, pandas operations, grouping/aggregating data, statistical tests (specifically: ANOVA, Post Hoc Analysis, and Mann Whitney U test), matplotlib (and seaborn) graphing, ML Classification, OneHot encoding, and rebalancing data.

Findings

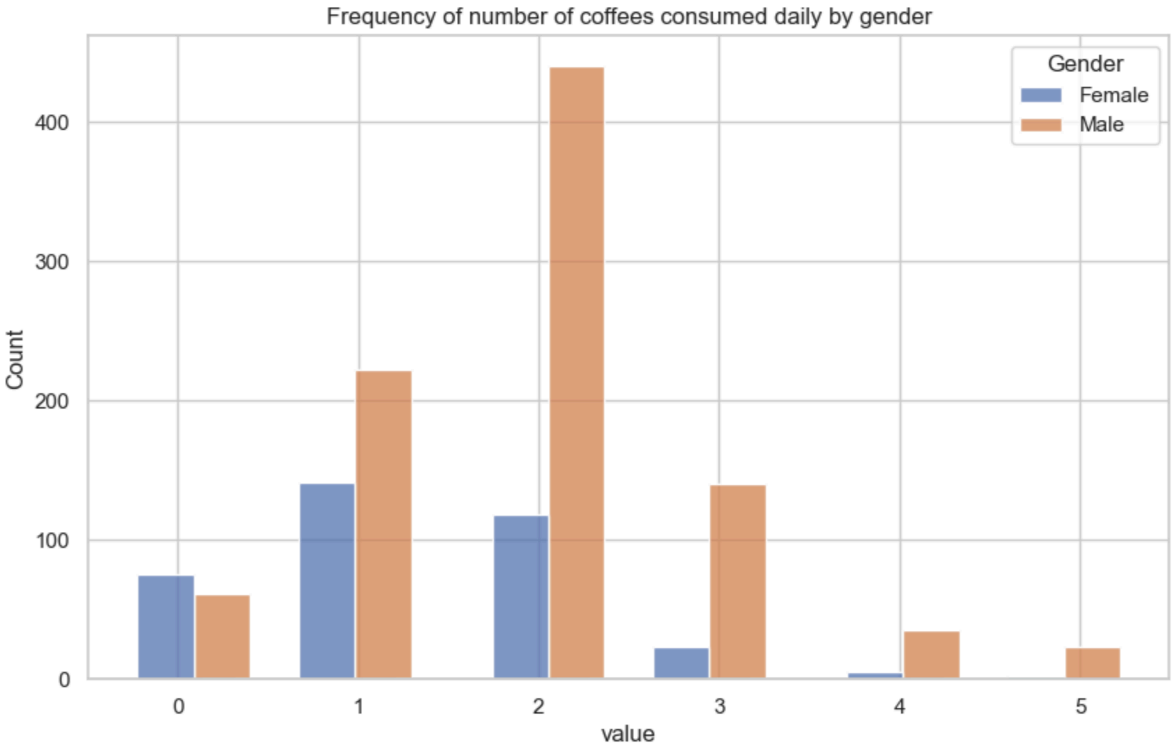
First I looked at who the biggest coffee buyers are. I decided that for this it would be interesting to see this in terms of age groups and gender. For age groups I had to remove a few groups due to lack of samples (<40). This histogram displays a count of how many people drink <1, 2, 3, 4, >5 cups of coffee per day by age group.



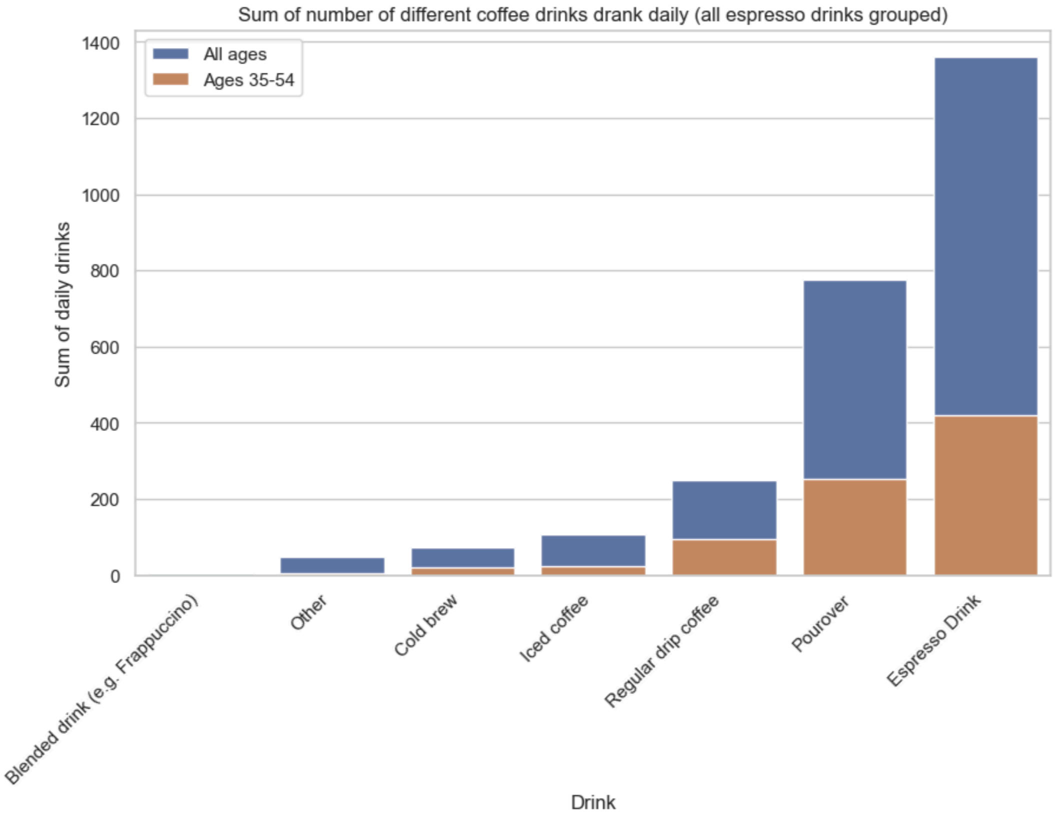
In this graph we can see that the data is *relatively normal* so I performed an ANOVA and Post Hoc analysis. The results of the analysis determined that age does make a difference in how much coffee someone drinks, **ages 18-34 would typically drink less coffee per day than ages 35-54.**



I did the same with genders but with the Mann Whitney U test (specifically male and female because I wanted to do a different test than the ANOVA again). My results were that **the typical amount of coffee drank by males and females are different**. Although we technically cannot conclude that males drink more coffee, it might be a safe assumption from the graph.

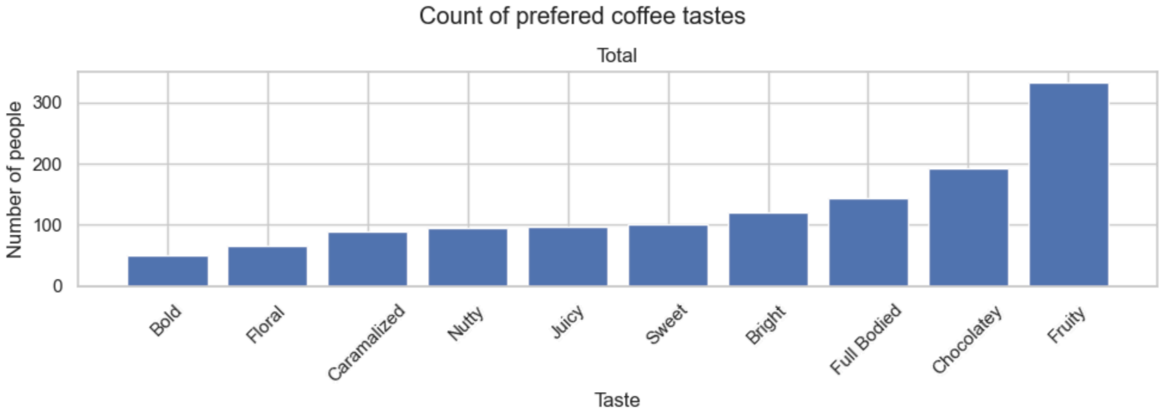


Secondly, I wanted to look at what kinds of drinks are popular so that we can determine what equipment might be needed. For this I grouped all the espresso drinks as one because you would need an espresso machine for each of them. I also added the age group 35-54 to see how the total compares to the age group that drinks the most coffee.

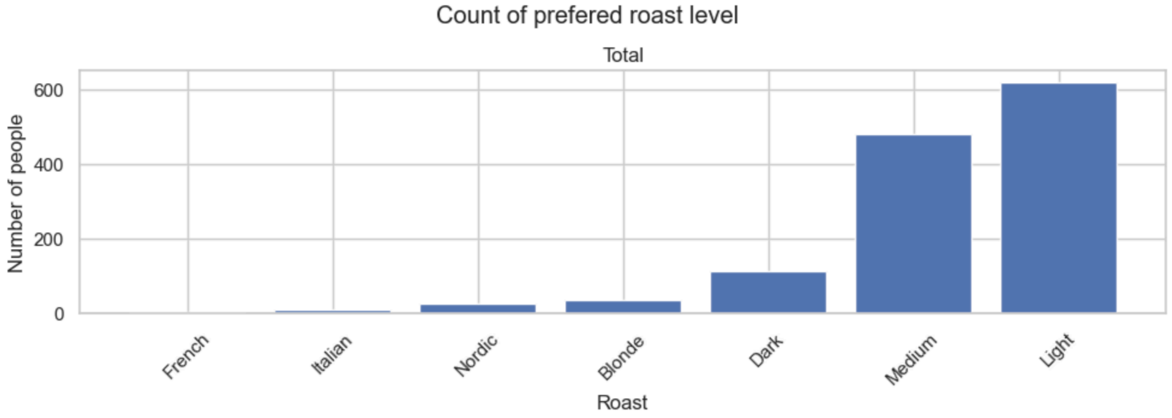


From this graph we can tell that an **espresso machine and pourover coffee maker are essential, whereas a blender might not be. This also tells us that hot drinks are preferred over cold drinks.**

Looking at flavours and roast levels can give us an insight into what types of beans we should buy.

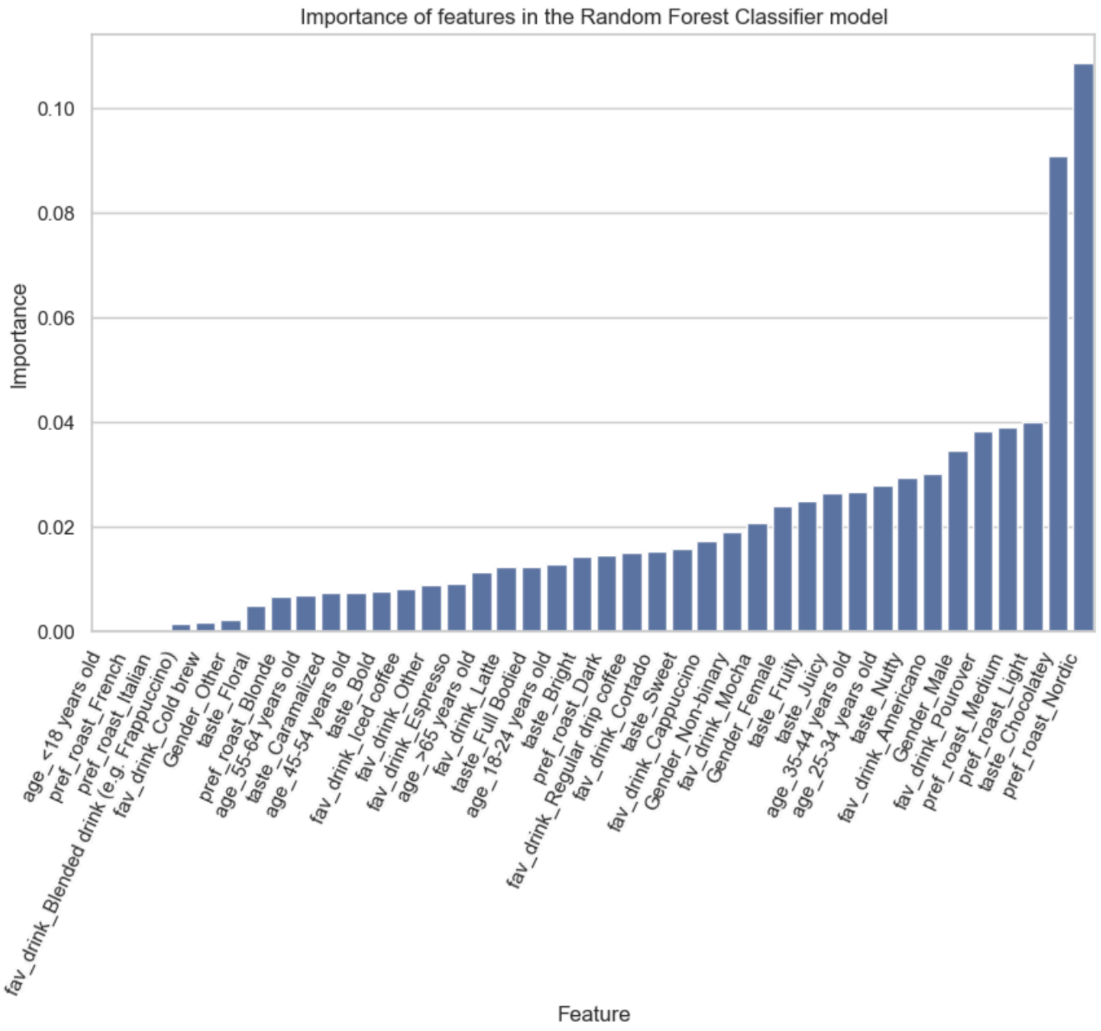


This graph tells us that we need to find **beans that are fruity, chocolatey, and/or full bodied.** Interestingly these preferences change between males and Females, with **males preferring fruity, full bodied, and bright**, and **females preferring chocolatey, fruity, and sweet.**



With this graph we can see that **in general people prefer light and medium roasts. With men preferring light roasts, and women preferring medium roasts.**

Finally, I wanted to use machine learning to identify big spenders vs small spenders. Big spenders are classified by people who typically spend more than $8 on one coffee and small spenders who spend less than $8. The features I chose to use are: gender, age group, preferred drink, preferred taste, preferred roast, milk options, and sweetener options. The model is relatively unsuccessful, it only has a testing and validation score of ~0.65 and ~0.60 respectively. This means that **these features are only slightly significant to determine whether someone spends a lot of money on coffee**. Another potential reason the scores are so poor is that there is not enough data. Either way the model is slightly successful so I graphed the feature importance.



It's important to note that this graph slightly changes with each run due to different testing/validation splits (and sampling), but the features are generally around the same area. This graph also does not tell us if the feature is important for determining if the individual is a big or small spender, so some further analysis is required for that. I did take a look at the Nordic roast and it seems that an overwhelming majority of people that prefer it are high spenders (~18 to 1). This is also slightly misleading because there are so few people who like Nordic roast to begin with. Further analysis might have to rebalance around this or figure out another way to draw more meaningful results via feature engineering.

Limitations

This data is pretty limited and biased since it comes from a youtuber and is more representative of his audience rather than the general population of the United States. Furthermore, there isn't enough data and diversity within the data, as there is hardly anyone younger than 18 or older than 55, and not many samples with genders that differ from male and female. More representation within the data could help us better understand our original questions.

If I had more time I would probably do more analysis on the ML model to determine which of the features are helping the most and the least to change them and make the model more accurate. I would also have loved to compare this data with some data from a real cafe or some other coffee survey data (I looked at the National Coffee Association 2025 survey data but it costs $1500 to use it).

Lastly, this report is only a summary of the most important findings throughout my analysis. There are a few tests and figure that are part of the notebooks but not part of the report so feel free to have a look.